ICA 2016 BUENOS AIRES
Buenos Aires – 5 to 9 September 2016
Acoustics for the 21st Century...

PROCEEDINGS of the 22nd International Congress on Acoustics

Signal Processing in Acoustics: Paper 608

Effects of acoustic degradations on cover song recognition

Julien Osmalskyj[(a)], **Jean-Jacques Embrechts**[(b)]

[(a)]University of Liège, Belgium, josmalsky@ulg.ac.be
[(b)]University of Liège, Belgium, jjembrechts@ulg.ac.be

Abstract:

Cover song identification systems deal with the problem of identifying different versions of an audio query in a reference database. Such systems involve the computation of pairwise similarity scores between a query and all the tracks of a database. The usual way of evaluating such systems is to use a set of audio queries, extract features from them, and compare them to other tracks in the database to report diverse statistics. Databases in such research are usually designed in a controlled environment, with relatively clean audio signals. However, in real life conditions, audio signals can be seriously modified due to acoustic degradations. For example, depending on the context, audio can be modified by room reverberation, or by added hands clapping noise in a live concert, etc. In this paper, we study how environmental audio degradations affect the performance of several state-of-the-art cover song identification systems. In particular, we study how reverberation, ambient noise and distortion affect the performance of the systems. We further investigate the effect of recording or playing music through a smartphone for music recognition. To achieve this, we use an audio degradation toolbox to degrade the set of queries to be evaluated. We propose a comparison of the performance achieved with cover song identification systems based on several harmonic and timbre features under ideal and noisy conditions. We demonstrate that the performance depends strongly on the degradation method applied to the source, and we quantify the performance using multiple statistics.

Keywords: Music recognition, cover songs, audio degradation, music information retrieval.

Effects of acoustic degradations on cover song identification systems

1 Introduction

Recent years have seen an increasing interest in Music Information Retrieval (MIR) problems. Such problems cover a wide range of research topics, such as automatic musical genre recognition, audio music transcription, music recognition, music recommendation, etc. In this paper, we address the problem of Cover Song Identification (CSI). CSI systems deal with the problem of retrieving different versions of a known audio query, where a version can be described as a new performance or recording of a previously recorded track [10]. Designing such systems is a challenging task because different versions of the same performance can differ in terms of tempo, melody, pitch, instrumentation or singing style. It is therefore necessary to design audio features and retrieval algorithms that are robust against changes in such characteristics. Most of existing works in the field of CSI compute pairwise comparisons between a query and a set of reference tracks in a search database [3, 6, 5]. To achieve that, audio features, usually corresponding to musical characteristics, are extracted from the audio signals. Audio features cover a wide range of musical characteristics such as the melody, the harmony (chords), the timbre, the tempo, etc. Once the features are extracted, a retrieval algorithm is used to compute similarity scores between the query and the tracks of the database. The goal of such an algorithm is to rank different versions of the query at the top of the returned list of tracks. The performance of a CSI system therefore depends on a trade-off between the selected audio features, and the retrieval algorithm. Most existing systems consider chroma features [4] as their main audio feature. Chroma features encode harmonic information in a 12-dimensional feature vector. Chroma vectors have been extensively used in the literature as they are robust against changes in the aforementioned musical characteristics. Ellis *et al.* [3] performs two dimensional cross-correlations of entire chroma sequences to highlight similar parts of the songs. Bertin-Mahieux *et al.* [1] consider the 2D Fourier transform magnitude coefficients of chroma patches to design a fast low-dimensional feature. Serra *et al.* [11] consider the entire chroma sequences of both tracks to be compared and use an alignment algorithm to compute a similarity score. Some authors also consider timbre features for CSI. In the work of Tralie *et al.* [13], the authors take into account the relative evolution of timbre to compute a similarity score. A comprehensive review of existing systems can be found in [8].

While many existing systems report a decent performance for CSI, they were evaluated in a controlled environment, usually with a single evaluation database. In this paper, we consider a selection of four existing systems and study the robustness of the features and the retrieval algorithms against acoustic degradations such as adding ambient noise at different levels, adding reverberation, simulating a live recording situation, applying harmonic distortion and convolving the query by the impulse responses of a smartphone microphone and speaker. Such experiments give us some information about how an existing CSI system would perform in real conditions, for example at a live concert, with a smartphone in a crowded room. To the best of our knowledge, we are the first to perform such a study for CSI. The results show that the studied systems are quite robust against audio degradations.

2 Studied cover song identification systems

We selected four state-of-the-art CSI systems for our study. This section describes briefly the selected systems. We refer the reader to the original works for detailled explanations.

2.1 Cross-correlation of chroma sequences (XCorr)

In that method, proposed by Ellis *et al.* [3], songs are represented by beat-synchronous chroma matrices. A beat tracker is first used to identify the beats time, and chroma features are extracted at each beat moment. This allows to have a tempo-independent representation of the music. Songs are compared by cross-correlating entire chroma-by-beat matrices. Sharp peaks in the resulting signal indicate a good alignment between the tracks. The input chroma matrices are further high-pass filtered along time. The final score between two songs is computed as the reciprocal of the peak value of the cross-correlated signal.

2.2 2D Fourier transform magnitude coefficients (2D-FTM)

In their work, Bertin-Mahieux *et al.* [1] split the songs into windows of 75 consecutive beat-synchronous chroma vectors, with a hop size of 1. 2D FFT magnitude coefficients are computed for each window, then are stacked together. A single 75x12 window is then computed as the pointwise median of all stacked windows. The resulting 900-dimensional patch is then projected on a 50 dimensional PCA subspace and the tracks are compared using the euclidean distance. This is one of the fastest feature available because it only computes 50-dimensional Euclidean distances, which is a straightforward operation.

2.3 QMax alignment of chroma sequences (QMax)

In Serra's *et al.* research [11], the authors first extract chroma features from both songs and transpose one song to the tonality of the other by means of the Optimal Transposition Index (OTI) method [9]. Then they form representations of the songs by embedding consecutive chroma vectors in windows of fixed length m, with a hop-size τ. Next they build a cross-recurrence plot (CRP) of both songs and use the QMax algorithm to extract features that are sensitive to cover song characteristics and update a similarity score.

2.4 Smith-Waterman alignment of timbre sequences (MFCC SW)

Tralie *et al.* [13] consider the use of timbre features rather than chroma features for cover song identification. They design features based on self-similarity matrices of MFCC coefficients and use the Smith-Waterman alignment algorithm to build a similarity score between two songs. Note that in contrast with other work considering MFCC features, they innovate by examining *relative* shape of the timbre coefficients. They demonstrate that using such features, cover song identification is still possible, even if the pitch is blurred and obscured.

3 Audio degradations

In this paper, we study how audio degradations affect the performance of four CSI systems. We selected six modifications to apply to the audio queries: add ambient restaurant noise, apply harmonic distortion, live recording simulation, convolution with the impulse response (IR) of a large hall, and the IRs of a smartphone speaker and a smartphone microphone. We used the Audio Degradation Toolbox (ADT) by Mauch *et al.* [7] to modify the audio signals. The ADT provides Matlab scripts that emulate a wide range of degradation types. The toolbox proposes 14 degradation units that can be chained to create more complex degradations.

3.1 Single degradations

We first apply non-parametric degradations. These audio modifications include: a live recording simulation, adding reverberation to the queries and convolving the queries by the IRs of a smartphone speaker and microphone. The live recording unit convolves the signal by the IR of a large room ('GreatHall', $RT_{30}@500Hz = 2s$, taken from [12]) and adds some light pink noise. The reverberation corresponds to the same convolution, without the added pink noise. The smartphone playback and recording simulations correspond to convolving the signal with the IR of respectively a smartphone speaker ('Google Nexus One') and the IR of the microphone of the same smartphone. The speaker has a high-pass characteristic and a cutoff at around 500Hz [7].

3.2 Parametric degradations

We add some ambient noise and distortion to the audio signals. The ambient noise corresponds to a recording of people in a pub. The recording is provided with the ADT [7]. We successively add the ambient noise at multiple SNR levels, from 30 dB to 5 dB to study how robust the systems are. We also successively add some harmonic distortion. To achieve this, the ADT applies a quadratic distortion to the signal. We iteratively applied the distortion with 2, 4, 6, 8 and up to 10 iterations. One iteration of quadratic distortion is applied as follows: $x = sin(x * \pi/2)$.

4 Experimental setup

4.1 Evaluation database

We evaluate our experiments on the Cover80[1] dataset [2]. The dataset contains 80 songs for which two versions are available, thus proposing a total of 160 tracks. While this is definitely not a large scale dataset, it has the advantage of providing audio data, allowing us to extract features straight from the audio. Other bigger datasets such as the Million Song Dataset[2] (MSD) or the Second Hand Song Dataset (SHS) are available, but they do not provide audio data. Rather than that, they provide pre-computed audio features that can be exploited in MIR

[1] http://labrosa.ee.columbia.edu/projects/coversongs/covers80/
[2] http://labrosa.ee.columbia.edu/millionsong

algorithms. For this specific research, we need the audio data so that we can modify the signals with respect to each degradation. We created 4 copies of the dataset for the single degradations (convolutions) and applied the convolutions with the default parameters provided by the ADT. For the ambient noise degradation, we created 5 additional copies, with added noise at SNRs of respectively 30 dB, 20 dB, 15 dB, 10 dB and 5 dB. For the distortion, we also created 5 copies, applying the distortion as explained in Section 3.2.

4.2 Features extraction

To use the four selected CSI systems, we need chroma features as well as MFCC features for the timbre. We extracted chromas from the audio using the Essentia library[3] with the HPCP [4] algorithm. Each chroma is elevated to a power of 2 to highlight the main bins, and then normalized to unit-norm. We first extracted beats location using a beat-tracker provided in the library, and computed 12-dimensional chroma features at each beat instant, with a sampling rate of 44.1kHz. For the computation of the self-similarity matrices based on MFCC features (see Section 2.4), we used some code that was kindly provided by the authors. The code makes use of the librosa[4] library to extract 20-dimensional beat-synchronous MFCC vectors.

5 Results

5.1 Evaluation methodology and metrics

For each modification of the database, we apply the same evaluation methodology. We consider all tracks of a noisy database (160 tracks) and we compare them to all tracks in the original database. Note that both databases contain exactly the same tracks. Each track in the noisy database is taken as a query and compared to 159 songs in the original database (we do not compare the query to itself). Using the similarity scores, we build an ordered ranking of 159 candidates for each query (highest score is considered most similar). We then look in the ordered ranking where the second version of each query is located (in terms of absolute position). We report the results in terms of Mean Rank (MR), which corresponds to the mean position of the second version (lower is better), in terms of Mean Reciprocal Rank (MRR) which corresponds to the average of the reciprocal of the rank of the identified version (higher is better). We also reports the proportion of queries for which the second version was identified at the first position (TOP-1), or in the 10 first returned tracks (Top-10).

5.2 Single degradations

Figure 1 compares the performance of the four selected CSI systems with respect to single audio degradations. The first column (blue) always corresponds to the performance of the system with no degradation. As one can observe on the figure, the degradation that affects the most each system is the *smartphone playback*. In particular, the 2D-FTM system has a significant loss of performance, with a decrease of 80% in terms of MRR.

[3] `http://essentia.upf.edu/`
[4] `https://github.com/librosa/librosa`

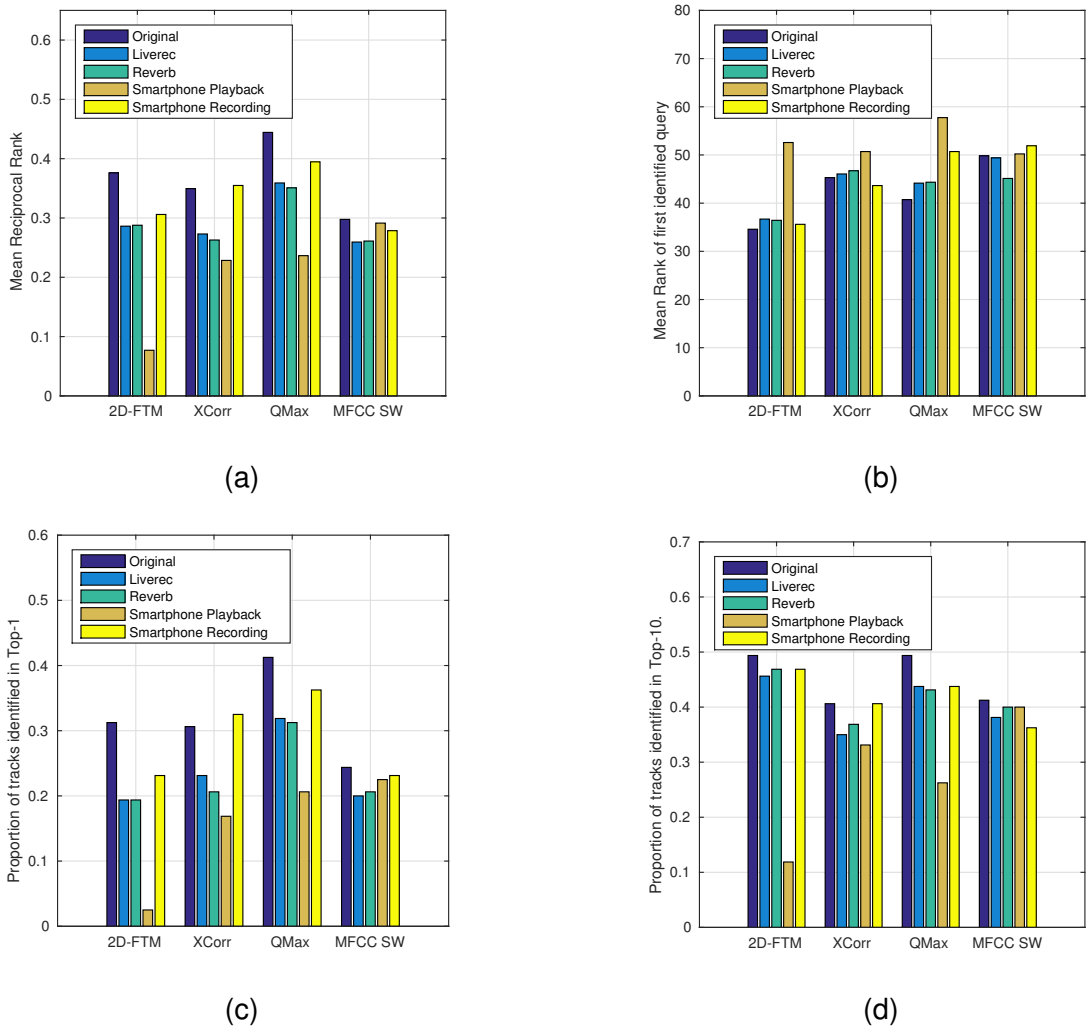

Figure 1: Evolution of the Mean Reciprocal Rank (a), the Mean Rank (b), the proportion of tracks identified in the Top-1 (c) and the proportion of tracks identified in the Top-10 (d) for single non parametric degradations.

This can be explained by the fact that the smartphone speaker has a high-pass characteristic with a cutoff at around 500Hz. Therefore, the spectrograms upon which the chromas are built lose much information compared to no degradation at all. Note that the timbre based system (MFCC SW) is definitely robust against the smartphone playback degradation. For both live recording simulation and added reverberation, all systems are not degraded significantly and performs similarly for both degradations. The most stable feature with respect to all degradations is the MFCC SW, with a maximum decrease of 13% in terms of MRR for the live recording simulation.

5.3 Ambient noise and distortion

Figure 2 shows the evolution of the performance of the four selected CSI systems when the percentage of ambient noise is increased (the SNR gets lower). We plot the results in terms of percentage of Noise-to-Signal amplitude ratio (NSR) to be able to represent the original point, with no noise added at all. We compute the NSR as follows:

$$NSR = \frac{100}{10^{\frac{SNR}{20}}} \tag{1}$$

We add the ambient noise with a decreasing SNR (resp. increasing NSR) at values of 30 dB, 20 dB, 15 dB, 10 dB and 5 dB.

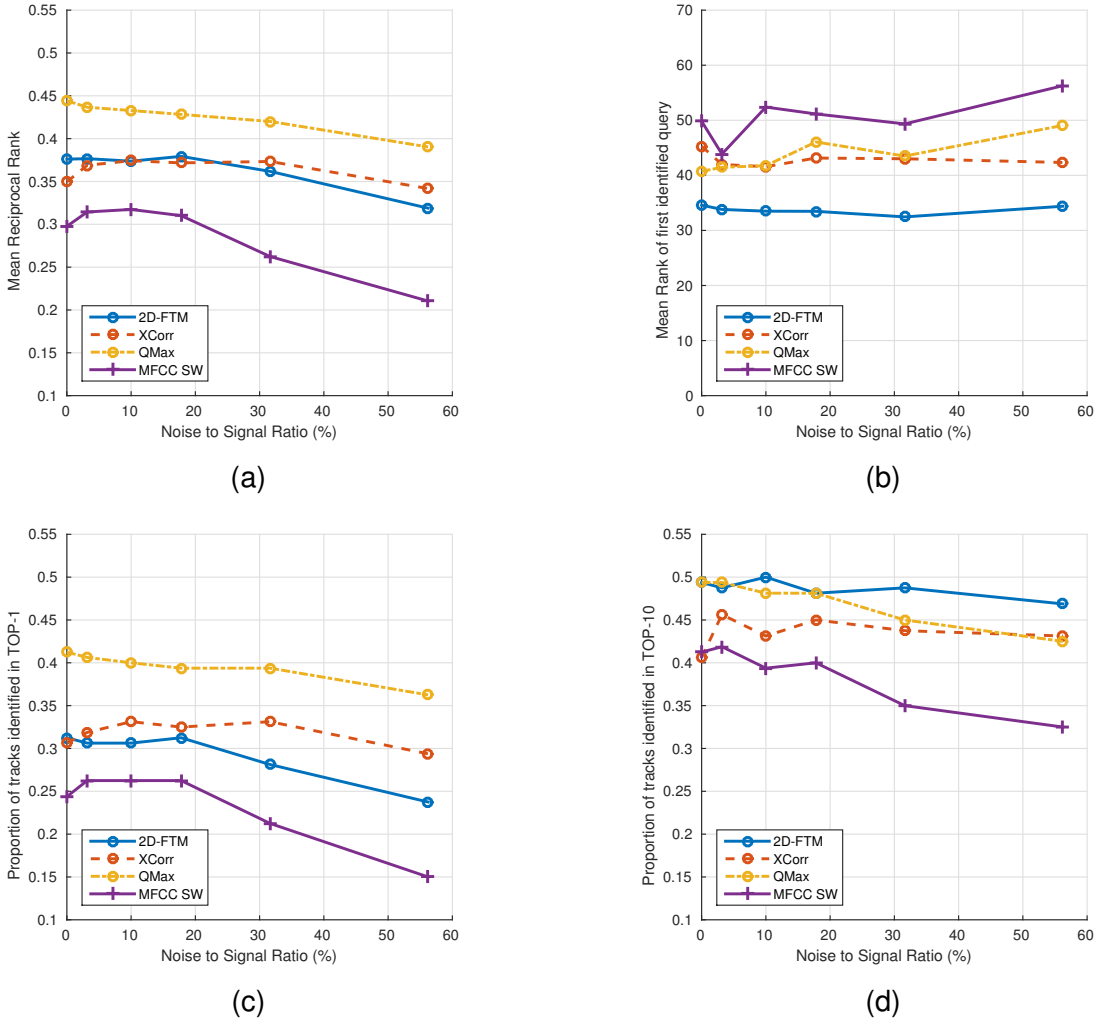

Figure 2: Evolution of the Mean Reciprocal Rank (a), the Mean Rank (b), the proportion of tracks identified in the Top-1 (c) and the proportion of tracks identified in the Top-10 (d) for an increasing ambient noise.

Adding an ambient noise to the original audio signal generates new frequencies in the spectrum. As the chroma features are computed based on that spectrum, we expect the performance to drop at some point. When adding up to 20% (SNR $\simeq 15$dB) of noise to the songs, all systems stay stable, with almost no loss in performance. Above 20%, the 2D-FTM and MFCC methods start to decrease the performance in terms of MRR, Top-1, and Top-10. In terms of MR, all methods stay stable at all noise levels. Note how the MRR and the Top-1 metrics render similar shapes. As both metrics take into account the position of the first match to the query, they seem to be strongly correlated.

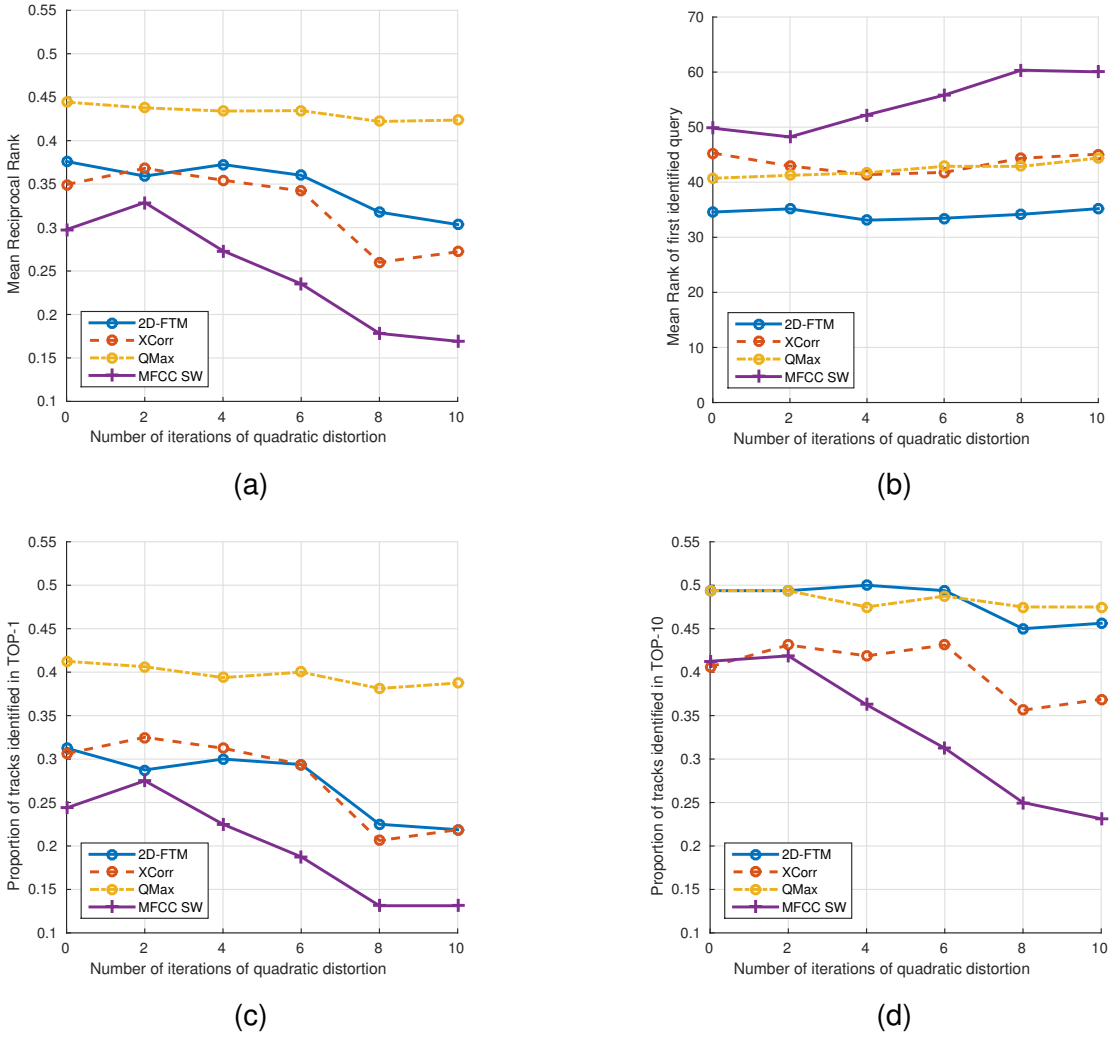

Figure 3: Evolution of the Mean Reciprocal Rank (a), the Mean Rank (b), the proportion of tracks identified in the Top-1 (c) and the proportion of tracks identified in the Top-10 (d) for an increasing number of iterations of quadratic distortion.

Figure 3 shows the evolution of the performance when we increase the number of iterations of quadratic distortion. The first observation one can make is that the QMax method is robust against any level of distortion, with respect to each metric. There is almost no loss of performance for the method. In terms of MRR and MR, 2D-FTM is also stable and does not decrease in performance. After two iterations, the MFCC method starts to drop in performance for each metric. This makes sense as the timbre is computed based on the harmonics of the signal. Applying quadratic distortion adds harmonics which can blur the timbre of the audio signal. The XCorr method drops in terms of MRR, Top-1 and Top-10 after 6 iterations, which makes it more robust than the MFCC method. Note that 6 iterations of distortion is clearly audible in the audio tracks, and the perceived music is strongly degraded compared to the clean song. In light of this, we can consider that all methods are pretty robust up to 4 iterations of quadratic distortion.

6 Conclusion

In this paper, we evaluated multiple state-of-the-art cover song identification systems with respect to several audio degradations. We first selected three methods based on chroma features, thus considering the harmonic content of the songs as the main feature. These methods use different retrieval algorithms to find cover songs in a reference database. We also chose a fourth method based on timbre feature rather than chroma features. The latter makes use of a sequence alignment algorithm to find relevant cover songs. We selected the Cover80 dataset for our research, and used the Audio Degradation Toolbox to perform a series of degradations of the database. We selected six degradations, corresponding to potential real-life modifications of the sound. The degradations include a live recording simulation, adding reverberation, convolving with the impulse responses of a smartphone speaker and microphone, adding a restaurant ambient noise at multiple levels and finally adding multiple iterations of quadratic distortion.

Overall, the methods based on chroma features perform similarly against all degradations. The worst performance is achieved through a convolution with a smartphone speaker and is produced by the 2D-FTM method. Convolving the signal by the microphone of the smartphone, however, does not degrade the performance significantly. Same goes for the live recording simulation and added reverberation. The timbre based method is extremely stable with single degradations, with almost no loss in performance with respect to all metrics, which makes it a robust method, although it performs worse than chroma based methods in a clean situation.

When adding ambient noise to the songs, all systems are stable up to 20% of added noise. After that limit, the timbre method decreases significantly, while the chroma based methods stay stable. When adding quadratic distortion, all systems but the timbre one stay stable up to 6 iterations. The MFCC based system drops after two iterations. After 6 iterations, XCorr and 2D-FTM lose some performance, but not significantly (less than 10% in terms of all metrics).

Overall, the studied systems can be considered stable against the applied audio degradations. We voluntarily degraded the signals significantly to push the limits of the systems, and the performance stays pretty stable. Future work involves analysing other cover song systems, and combining them together to study how the robustness against audio degradations performs.

References

[1] T. Bertin-mahieux and D. Ellis. Large-scale cover song recognition using the 2d fourier transform magnitude. In *Proceedings of the 13th International Society for Music Information Retrieval (ISMIR)*, pages 241–246, 2012.

[2] D. Ellis and C. Cotton. The 2007 labrosa cover song detection system. *Mirex 2007*, 2007.

[3] D. Ellis and G. Poliner. Identifying "Cover Songs" withchroma featires and dynamic beat tracking. In IEEE, editor, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–16, New York, 2007. IEEE.

[4] E. Gómez. Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing*, 18(3):294–304, 2006.

[5] E. Humphrey, O. Nieto, and J. Bello. Data Driven and Discriminative Projections for Large-scale Cover Song Identification. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 4–9, 2013.

[6] M. Khadkevich and M. Omologo. Large-Scale Cover Song Identification Using Chord Profiles. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 5–10, 2013.

[7] M. Mauch and S. Ewert. The Audio Degradation Toolbox and Its Application To Robustness Evaluation. In *14th International Society for Music Information Retrieval Conference (ISMIR)*, pages 2–7, Curitiba, Brazil, 2013.

[8] J. Serrà. *Identification of Versions of the Same Musical Composition by Processing Audio Descriptions*. PhD thesis, 2011.

[9] J. Serrà and E. Gómez. Transposing Chroma Representations to a Common Key. In *IEEE Conference on The Use of Symbols to Represent Music and Multimedia Objects*, pages 45–48, 2008.

[10] J. Serrà, E. Gomez, P. Herrera, and X. Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech and Language Processing*, 16(6):1138–1151, 2008.

[11] J. Serrà, X. Serra, and R. G. Andrzejak. Cross recurrence quantification for cover song identification. *New Journal of Physics*, 11(9), 2009.

[12] R. Stewart and M. Sandler. Database of omnidirectional and B-format room impulse responses. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 165–168, Dallas, USA, 2010.

[13] C. Tralie and P. Bendich. Cover song identification with timbral shape sequences. In *16th International Society for Music Information Retrieval Conference (ISMIR)*, pages 38–44, Malaga, 2015.